#### PATENT ABSTRACTS OF JAPAN

(11)Publication number:

2001-318766

(43) Date of publication of application: 16.11.2001

(51)Int.CI.

G06F 3/06 G06F 12/08 G06F 12/16 G11B 19/02

(21)Application number: 2000-135013

(71)Applicant: NEC SOFTWARE SHIKOKU LTD

(22)Date of filing:

08.05.2000

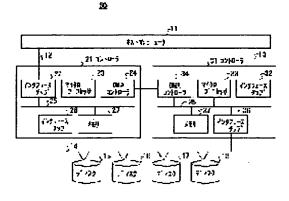
(72)Inventor: SHIRAISHI KAZUYA

#### (54) DISK ARRAY DEVICE AND CACHE MEMORY CONTROL METHOD

#### (57)Abstract:

PROBLEM TO BE SOLVED: To provide a disk array device and a cache memory control method for executing second and succeeding pieces of write command processing without putting loads on the processor of a main controller.

SOLUTION: This disk array device is composed of two controllers and executes the double write of write data in the two controllers without using the hardware of a shared memory. The write data are held in both of the two controllers and a means is provided for making the write data held in the other one of the two controllers mutually referable when a fault is generated in one of the controllers.



#### **LEGAL STATUS**

[Date of request for examination]

17.04.2001

[Date of sending the examiner's decision of

12.02.2003

rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

### Best Available Copy

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

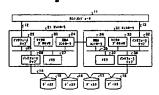
(12) 公開特許公報(A) (19)日本国特許庁 (JP) (11)转許出願公園番刊 特別2001-318766 (P2001-318766A) (43)公開日 平成13年11月16日(2001.11.16) 銀列配号 304 540 5-70-)\*(##) 5 B 0 0 5 5 B 0 1 8 (51) Int.Cl.' G 0 6 F 304E GOSF 511Z 541C 12/0 12/08 6 B 0 6 5 南京項の数14 OL (全 12 月) (21)出職書号 00 - 135013( P2000 -- 135013) (71) 出版人 000180379 四国日本電気ソフトウェア株式会社 (22)出版日 平成12年5月8日(2000.5.8) 爱健県松山市农山 4 丁目760番地 白石 和也 今四年份山市太山4丁自760 四国日本省 (72) 発明者 弁智士 頌 城之 Fターム(参考) SB005 JJ01 JJ11 MHII NNO2 990-58018 CAD4 HAD2 HA35 KAD3 NA14 58086 BAD1 CA12 CA30 CC08 CE12

(54) 【発明の名称】 ディスクアレイ装置およびキャッシュメモリ制御方法

#### (57) 【要約】

【練題】 本発明は、2回目以降のライトコマンド処理 においてメイトコントローラのプロセッサに負荷をかけ ることなく実行できるようになるディスクアレイ装置お よびキャッシュメモリ制御方法を提供することを課題と

・ 【解決手段】 2つのコントローラで構成されかつ共有 メモリのハードウェアを用いることなく当故2つのコン トローラでライトデータの2重響きを実行するディスク アレイ装置であって、ライトデータを斡記2つのコント ローラの両方に保持するとともに、前記コントローラの 一方に故障が発生した際に前記2つのコントローラの他 方に保持されているライトデータを相互参照可能とする



12. 13. 14- HR-CA

2003 04 24 10.41

### ディスクアレイ装置およびキャッシュメモリ制 都方法

特開2001-318766

【緯求項12】 前記ピットマップは、各セグメントに 対応するキャッシュのどの位置にな効なデータが存在す るかを示すマップであることを特徴とする緯収項()に 紀載のキャッシュメモリ制御方法。

【静泉項13】 前記ライトキャッシュは、各セグメントに対応するキャッシュメモリであることを特徴とする **請求項12に記載のキャッシュメモリ制御方法。** 

【翻求項14】 前記ホストコンピュータからライトコ マンドを受信した際に前記コントローラの一方は、自己 の前紀セグメント管理情報を参照して自身がシーズして いるセグメントにヒットしているかを確認する工程と、 ヒットしていなければ沢用パスを介して前紀2つのコン トローラの他方にシーズコマンドを送信してセグメント モシーズする工程と、救記ホストコンピュータからライ トデータを受信して自己の前起ライトキャッシュに格納 し当鉄格納位置に対応した自己の前記ピットマップを有 **効状性に更新する工程と、自己の点記ライトキャッシ**: のデータおよび自己の前記ピットマップを前記2つのコ ントローラの他方のDMAコントローラモ介して前記2 つのコントローラの他方側の前記ライトキャッシュおよ び前記ピットマップに送信してライトコマンドを受信す る工程と、前記コントローラの…ガがシーズしているセ グメントにヒットした場合に前紀2つのコントローラの 他方にシーズコマンドを送付することなく、前記ホスト イトキャッシュに格納し当鉄格納位置に対応した自己の 前配ビットマップを有効状態に更新する工程と、自己の 前配ライトキャッシュのデータおよび自己の前配ビット マップを向記2つのコントローラの他方のDMAコント ローラを介して前記2つのコントローラの他方の前記ラ イトキャッシュと前記ピットマップに追ばする 上程を存 することを特徴とする結束項13に記載のキャッシュ。

#### [発明の群組な説明]

【発明の興する技術分野】 4発明は、2コントローラ橋 成でかつ共有メモリのハードウェアを持たないディン アレイ装置のキャッシュメモリ制御技術に係り、特にメ イトコントローラのセグメントをシーズしてそのキャッ シュアドレスとビットマップを保持することにより、2 図目以降のライトコマンド処理においてメイトコントロ ーラのプロセッサに負債をかけることなく実行できるようになるディスクアレイ装置およびキャッシュメモリ動 参方法に関する.

【従来の技術】従来、ライトデータの2.00mのおお技術とし ては、例えば、キャッシュメモリを2つのコントローラ から独立させて設け当該両コントローラからアクセス可 魔な共有メモリ構成とする方式と、コントローラ内にキ 『ツシコメモリを配置し何らかの方法でキャッシュデ

タをメイトコントローラに送信する方式等が開示されて 116.

【0003】このような先行技術としては、例えば、料 期平10-222423号公翰(第1従来技術)に記載 のキャッシュメモリ制御方式がある。図5は、第1従来 技術のキャッシュメモリ制御方式を説明するための機能 プロック図である。図5を参照すると、第1従来技術の キャッシュメモリ制御方式は、プロセッサ100、プロセッサ200、共有メモリ300、および入出力制御装 置400が共有パス10に接続されたMPシステムであ

【0004】プロセッサ100は、ライトバック方式で 制御されるキャッシュメモリ 1 1 0 と、キャッシュメモ リ 1 1 0 に入出力されるアドレス信号線、データ信号

線、制御信号線などを制御するキャッシュメモリ制御回 路120と、キャッシュメモリ110のアドレス部と状 盤部の写しを格納するコピータグ 130と、コピータグ 130に入出力されるアドレス信号線、データ信号線、

制御信号線などを制御するコピータグ制御回路140 と、演算処理やプログラムの実行制御を行う演算制御回 路150と、演算制御回路150と共有パス10とのイ ンターフェースを制御するバスインターフェース回路! 60と、共有パス10上のアクセス要求を検出しコピー タグ制御回路140やパスインターフェース回路160

への指示を行うパス監視回路170とで構成される。 ャッシュメモリ110は、格納している各プロックの.b 位アドレスを記憶するアドレス部 I 1 1 と、各ブロック の状盤値を格納する状態部 I 1 2 と、各ブロックのデー 夕を記憶するデータ部113とから構成される。ここで

いう上位アドレスとは、アドレスをピット列としてみた 場合に、ブロック内のパイト位置を指定するピット列お よびキャッシュメモリ内のプロック位置を指定するピッ ト列を除いた上位ピット列を指す。アドレス選111と 状態器112とを合わせて、タグ部と呼ぶ。コピータグ

130は、キャッシュメモリ110のタグ部の写しを格 動するものであり、キャッシュメモリ110のアドレス 個111の写しを格納するアドレス郎131と、キャッシュメモリ110の状態部112の珍しを格納する状態 部132とから構成される。キャッシュメモリ110の 状態部 1 1 2 とコピータグ 1 3 0 の状態部 1 3 2 とはそれぞれ、ブロックごとに 4 つの状態値を格納することが

できる。4つの状態値とは、無効状態、共有状態、排他 一数状態、および排他変更状態である。キャッシュメモ り制御回路120は、キャッシュメモリ!10の任意の ブロックの鉄他変更状態への選称とそのブロック位置を コピータグ130に通知する信号を信号線121を通し

て出力する。

同様に構成されており、プロセッサで00の構成をあむ ・・ 50 それぞれプロセッサ 100の相当する構成要素と同様で 【特許請求の範囲】

「競求項」 2つのコントローラで構成されかつ共存 メモリのハードウェアを用いることなく当該2つのコン トローラでライトデータの2重書さを実行するディスク アレイ装置であって、

ライトデータを仰記2つのコントローラの両方に保持す るとともに、前起コントローラの一方に故障が発生した 摩に前記2つのコントローラの他方に保持されているラ イトデータを相互参照可能とする手段を存することを特 数とするディスクアレイ装置。 【請求項2】 ホストコンピュータとディスクとの間に

前記2つのコントローラの両方を有し、

前記ホストコンピュータと前記2つのコントローラの両 方が汎用パスで接続されるとともに、前記2つのコント ローラの両方と前紀ディスクが汎用パスで接続されてい ることを特徴とする緯求項1に記載のディスクアレイ装

【前求項3】 前記2つのコントローラのそれぞれは、 マイクロブロセッサと、リードまたはライトのデータを キャッシングしかつプロセッサの制御情報を記憶するメ モリと、前紀ホストコンピュータとの通信を制御する第 1のインタフェースチップと、前紀ディスクとの通信を 制御する第2のインタフェースチップと、仰起メモリの 任意アドレスのデータを削起2つのコントローラの他方 の前紀メモリの任意アドレスに送信できるDMAコント ローラを有していることを特徴とする請求項1または2 に紀載のディスクアレイ装置。

【趙沢項4】 セグメント単位でキャッシュを管理する 前記メモリには、セグメント管理情報と、ビットマップ と、ライトキャッシュが定義されていることを特徴とす

る静泉項3に配載のディスクアレイ装置。 【静泉項5】 - 蔚起セグメント管理情報は古セグメント の状態を保持するものであって、シーズの状態を示すシ ーズフラグと、ロジカルユニット番号と、ロジカルブロ ックアドレスと、自コントローラの両方のピットマップ 格納位置を示す自ピットマップアドレスと、自コントロ ラのキャッシュ位置を示す白キャッシュアドレスと 前紀2つのコントローラの他方のピットマップ格納位置 を示す他ピットマップアドレスと、 前紀2つのコントロ ーラの他方のキャッシュ位置を示す他キャッシュアドレ スを有することを特徴とする請求項4に記載のディスク

【類求項6】 前記ピットマップは、各セグメントに対 応するキャッシュのどの位置に行効なデータが存在する かを示すマップであることを特徴とする請求項5に配載 のディスクアレイ装置。

に対応するキャッシュメモリであることを特徴とする誰 **東瓜6に記載のディスクアレイ装置**。

【緯求収8】 前記ホストコンピュータからライトコマ 50

ンドを受信した際に前記コントローラの・・方は、自己の 和記セグメント管理機能を参照して自身がシーズとでいるセグメント管理機能を参照して自身がシーズとでいるセグメントにヒットしているかを確認し、ヒットしていなければ汎用パスを介して利起2つのコントローラの

他力にシーズコマンドを送信してセグメントをシーズ し、前記ホストコンピュータからライトデータを受けし て自己の前記ライトキャッシュに格納し当故格納位置に 対応した自己の前記ピットマップを有効状態に更新し、

自己の前記ライトキャッシュのデータおよび自己の何記 ビットマップを前記2つのコントローラの他方の前記D ルイン・ローラを介して前記2つのコントローラの他 方頼の前記ライトキャッシュおよび前記ビットマップに 送信してライトコマンドを受はし、前記コントローラの 一方がシーズしているセグメントにヒットした場合に約 記2つのコントローラの他方にシーズコマンドを送信す

ることなく、前記ホストコンピュータからライトデータ を受信して自己の前記ライトキャッシュに格納し当該格 納位置に対応した自己の点記ピットマップを有効状態に **更新し、自己の前記ライトキャッシュのデータおよび**自 己の前紀ピットマップを前記2つのコントローラの他方

の約記DMAコントローラを介して前記2つのコントローラの地方の前記ライトキャッシュと前記ピットマップ に送ばすることを特徴とする精栄項?に記載のディスク

【錆求項9】 2つのコントローラで構成されかつ共作 メモリのハードウェアを用いることなく当故2つのコン ・うでライトデータの2重要きを実行するディスク アレイ装置に対して、ライトデータを前紀2つのコント ローラの両方に保持するとともに、前記コントローラの 一方に故障が発生した際に前記2つのコントローラの也 方に保持されているライトデータを相互参照可能とする C程を有することを特徴とするキャッシュメモリ制御方

【緯求項 1 0】 前記 2 つのコントローラのそれぞれに 設けられリードまたはライトのデータをセグメント単位 でキャッシングしかつプロセッサの知識情報を記憶する ・ エファッシュー・ ---(モリに、セグメント管理情報と、ピットマップと イトキャッシュを定義する工程を有することを特徴とす る請求項9に記載のキャッシュメモリ制御方法。 【請求項11】 前記セグメント管理情報は各セグメン

トの状態を保持するものであって、シーズの状態を示す シーズフラグと、ロジカルユニット番号と、ロジカルブ ロックアドレスと、自コントローラの両方のピットマッ ブ格納位置を示す自ビットマップアドレスと、自コント ローラのキャッシュ位置を示す自キャッシュアドレス

と、前記2つのコントローラの他方のピットマップ格納 位置を示す他ピットマップアドレスと、前記2つのコン トローラの処方のキャッシュ位置を示す他キャッシュブ ドレスを有することを特徴とする諸求項10に記載の主 ャッシュメモリ制御方法。

### ディスクアレイ装置およびキャッシュメモリ制 和方法

特開2001-318766

ある。すなわち、プロセッサ200は、アドレス部21 1、状態期212、およびデータ第213を備えるキャ ・、いまはと・1と、こまじり フロと・3を構んしてイ ッシュメモリ210と、キャッシュメモリ対響回路22 0と、アドレス部23!および状態電232を備えるコ ピータグ230と、コピータグ制御回路240と、演算 制御回路250と、パスインターフェース回路260 と、バス監視回路270とで構成される。

【0006】共有メモリコロロは、共有バス10に接続 され、プロセッサ 100、プロセッサ200、および入 出力制御装置400から共有パス10上に要求された。 モリリードアクセスやメモリライトアクセスに対して応 答可能な記憶装置である。

【0007】入出力制御装置400は、特定の入出力装 置、例えば磁気ディスク装置(図示せず)とのインター フェースを持ち、共有パス 10を介して共有メモリ30 0にリード、プライト可能な装置である。

(0008) 共有パス10は、接続された各装置が相互 にアクセス可能なためのアドレスは月線、データ信号 線、その他リードライトコマンド信号線、パス間停はり 線、タイミングは号線などの制御信号線を含むバスであ る。共有バス10は、特に、リード要求に対する応答う ータがプロセッサにより変更されているか否かを通知する変更は号報§ 1 1 と、リード要求に対する応答データ が複数のプロセッサで共有されるか否かを通知する共有 付号線512とを含む。

【0009】また他の従来技術としては、例えば、特開 平6-67979月公報 (第2従来技術) に起献の主記 懲装置の制御方式がある。図6は、第2従来技術の主記 億装置の制御方式を説明するための機能ブロック図であ

【0010】図6を参照すると、第2従来技術の主記憶 装置は、現用系情報処理システム(以下SYSA)61 1と、障害時に処理を引き継ぐ待機情報処理システム (SYSB) 612とから成っている。SYSA611 はプロセッサ (以下CPU) | aと、CPU | aにアド レスパス5 a、データパス6 aで接続されたコピーパッ ク・キャッシュメモリ (以下CBCM) 2nと、CBC M2nにアドレスパス7n、データパス8nで接続され た主記憶装置(以下MEM) 3 a と、CPU 1 a にアド レスパス5a.データパス6aで接続されたコピーパ ・パッファ装置(以下CBBM) 4 a とを存して構成 している。

【0011】また、SYSB612はプロセッサ (以下 CPU) 15と、CPU15にアドレスパス55、デー タバス6Dで接続されたコピーバック・キャッシュメモ リ (以下CBCM) 2Dと、CBCM2Dにアドレスバ 3.7 b. データバス8 bで接続された主足機拡進 (EFF MEM) 3 bと、CPU | bにアドレスパスちも、デー タバスもしで接続されたコピーパック・パッファ装置 (以下でBBM) すりとを有して構成している。

[00121 SYSA6110CBBM4attSYSB 612のMEM3わにアドレスパスタッとデータパスト Oaとで接続され、SYSB6!2のCBBM4bはS YSA611のMEM3aにアドレスパス9bとデータ パス10日とで接続され、本適用例はSYSA611と SYSB612とが互いに対称形に接続されたシステム

構成となっている。 【0013】このように、第2従来技術では、コントロ ーう内にキャッシュメモリを配置し、ハードウェアによってメイトコントローラにライトデータを追信する仕組みとして、CPUからコピーパック・キャッシュメモリ (CBCM) に送信されるライトデータをCBBMで標 取りしてメイトコントローラに送信しているが、例え

は、PCIパス(Peripheral Component Interconnect bus:システムパス・ローカルパス)のような代用パスを用いる場合に おいては模取りという仕様は定義されていないため、内 部パスはオリジナルパスで構成することになる。

【0014】また第3の従来技術として、例えば、図7 に示すようなディスクアレイ装置が開示されている。図 7は、第3従来技術のディスクアレイ装置を説明するた めの機能プロック図である。すなわち、第3従来技術の ディスクアレイ装置は、ホストコンピュータ11とディ スク15. …. 18との間に2つのコントローラ21

3 1を存するものである。ホストコンピュータ1 1 とコ ントローラ21、31は、例えば、SCSI(スカジ ー:スモール・コンピュータ・システム・インタフェ-ス) パスのような汎用パス12、13で接続されてい

る。コントローラ21、31とディスク15、…、18 10 は、例えば、SCSIパスのような汎用パス14で接続 されている.

【0015】コントローラ21は、マイクロプロセッサ 23と、リードまたはライトのデータをキャッシングし かつプロセッサの調御情報を記憶するメモリ27と、ホ ストコンピュータ11との適位を制御するイン スチップ22と、ディスク15, …, 18との適似を制御するインタフェースチップ26を有し、コントローラ内の各チップは内部パス25で接続されている。

【0016】祠様に、コントローラ31は、マイクロブ ロセッサ33と、リードまたはライトのデータをキャッ シングしかつプロセッサの制御情報を記憶するメモリ3

7 と、ホストコンピュータ!」との通信を制御するイン タフェースチップ32と、ディスク15、…、18との タフェースチップ32と、ディスク15、…、18との 通信を制御するインタフェースチップ36を有し、コン ・ローラ内の各チップは内部パス35で接続されてい

[0017]

1 従来技術には、共介メモリ300は、コントローラの 放験交換によって一緒に取り外されないように、独立し た基板 (パッケージ) で作成されているため、当故メモ りの故障を考慮すると、共有メモリ300の2重化が必要になり、回路構成が大きくコストが高くなるという欠 点があり、中小型システムでは採用することが嬉しいと いう問題点があった。 【0018】また、上記第2従来技術は、内部パスをオ

リジナルパスで構成することになるため、汎用パスを採 用できる場合はインタフェースチップなどを汎用品の中 から自由に選択でき高性能な装備を安価に人手できる。 リットがあるものの、オリジナルバスではこのような鍵 成が難しいという問題点があった。

【0019】そして、上起第3従来技術は、2重審さに はディスク側の汎用パス14を使用するため、処理のオ ーパーヘッドが大きいという問題点があり、また。 トコントローラのプロセッサに負荷がかかるという問題

【0.0.2.0】本発明は斯かる問題点を個みてなされたも のであり、その目的とするところは、メイトコントロー うのセグメントをシーズしてそのキャッシュアドレスと ピットマップを保持することにより、2回日以降のライ 20 トコマンド処理においてメイトコントローラのプロセッ サに負債をかけることなく家行できるようになるディス クアレイ装置およびキャッシュメモリ制御方社を提供す る点にある。

【課題を解決するための手段】この発明の請求項1に記 最の発明の長旨は、2つのコントローラで構成されかつ 共有メモリのハードウェアを用いることなく当故2つの コントローラでライトデータの2電響さを実行するディ スクアレイ装置であって、ライトデータを向記2つのコ ントローラの両方に保持するとともに、何起コントロー **うの一方に故障が発生した際に前記2つのコントローラ** の他方に保持されているライトデータを相互参照可能と する手段を有することを特徴とするディスクアレイ装置 、たちする。また、この発明の雌栄項2に記載の発明の整 行は、ホストコンピュータとディスクとの間に前記2つ のコントローラの両方を有し、前記ホストコンピュータ と前記2つのコントローラの両方が利用パスで接続されるとともに、前記2つのコントローラの両方と前記ディ スクが汎用バスで接続されていることを特徴とする請求 項1に記載のディスクアレイ袋器に存する。また、この 発明の顧求項3に記載の発明の委員は、前記2つのコン トローラのそれぞれは、マイクロプロセッサと、リード またはライトのデータをキャッシングしかつプロセッサ の制御情報を記憶するメモリと、前記ホストコンピュー タとの通信を制御する第1のインタフェースチップと、 前紀ディスクとの通信を制御する第2のインタフェース チップと、前記メモリの任意アドレスのデータを前記2 つのコントローラの他方の前記メモリの任意アドレスに 透信できるDMAコントローラを有していることを特徴 50 スクアレイ装置に存する。また、この発明の請求項9に

とする顔以近1または2に記載のディスクアレイ装置に 存する。また、この発明の請求項4に記載の発明の委行は、セグメント単位でキャッシュを管理する前記メモリ には、セグメント管理情報と、ピットマップと、ライト

ッシュが定義されていることを特徴とする精求項3 に記載のディスクアレイ装置に存する。また、この発明 の請求項5に記載の発明の整旨は、前記セグメント管理 情報は各セグメントの状態を保持するものであって、シ ーズの状態を示すシーズフラグと、ロジカルユニット番 母と、ロジカルブロックアドレスと、自コントローラの

**両方のピットマップ格納位置を示す自ピットマップアト** レスと、自コントローラのキャッシュ位置を示す自キャッシュアドレスと、前記2つのコントローラの他方のビ ットマップ格納位置を示す他ピットマップアドレスと、

和記2つのコントローラの他方のキャッシュ位置を示す 他キャッシュアドレスを有することを特徴とする論求項 4 に紀載のディスクアレイ装置に存する。また、この発 明の讃求項 6 に記載の発明の要行は、前記ピットマップ は、各セグメントに対応するキャッシュのどの位置に有

効なデータが存在するかを示すマップであることを特徴 とする額求項5に記載のディスクアレイ装置に存する。 また、この発明の請求項でに記録の発明の委員は、前記 ライトキャッシュは、各セグメントに対応するキャッシュメモリであることを特徴とする請求項 6 に配戴のディ スクアレイ装置に存する。また、この発明の緯収項8に 紀載の発明の要旨は、前紀ホストコンピュータからライ

トコマンドを受信した際に前記コントローラの一方は、 自己の前記セグメント管理情報を参照して自身がシーズ しているセグメントにヒットしているかを確認し、ヒッ しているセクメントにピットしているかを確認していましたいいるでは、別用バスを入して前起で3のコントローラの他方にシーズコマンドを透信してセグメントをシーズし、前起ホストコンピュータからライトデックを 他して自己の前起ライトキッシュに接続け、当覧格時位 誰に対応した自己の前起ビットマップを有効収録に更新

銀に対応した自己の報記ピットマップを有効は壁に要所 に、自己の報記とイトキャッシュのデータおよび自己の 都記ピットマップを相配2つのコントローラの他方の報 記DMAコントローラを介して相配2つのコントローラの 他的場解の報題とゲートキッシュおよび解記とテーマップに連信してライトコマンドを受信し、個記コントロー ラの一方がシーズといるセグメントにヒットした場合 に都記2つのコントローラの他方にシーズフンドを追 はずることなく、報記はストコンピュータのライトデ

ータを受信して自己の前記ライトキャッシュに格納し当 被格納位置に対応した自己の前記ピットマップを有効状 撃に更新し、自己の前記ライトキャッシュのデータおよ び自己の前記ピットマップを前記2つのコントローラの 他方の前記DMAコントローラを介して前記2つのコン トローラの他方の前記ライトキャッシュと前記ピットマップに送信することを特徴とする請求項?に記載のディ

ディスクアレイ装置およびキャッシュメモリ制 和方法 紀載の発明の要旨は、2つのコントローラで構成された

に取りた明めな行い。2 かっといい。 ついれていない ひれんメモリのハードウェアを用いることなく当該2つのコントローラでライトデータの2 重選さを実行するディスクアレイ装置に対して、ライトデータを耐起2つの コントローラの両方に保持するとともに、前起コントローラの一方に故障が発生した際に前記2つのコントロー うの他方に保持されているライトデータを相互を限可能 とする工程を有することを特徴とするキャッシュメモリ 制御方法に存する。また、この発明の請求項10に記載 の発明の要旨は、前記2つのコントローラのそれぞれに 殺けられリードまたはライトのデータをセグメント単位 でキャッシングしかつプロセッサの制御情報を記憶する メモリに、セグメント管理情報と、ビットマップと、ライトキャッシュを定義する工程を有することを特徴とす る請求項9に記載のキャッシュメモリ制御方法に存す る。また、この発明の請求項11に記載の発明の報告 は、叔紀セグメント管理情報は各セグメントの状態を保 持するものであって、シーズの状盤を示すシーズフラグ と、ロジカルユニット番号と、ロジカルブロックアドレ スと、自コントローラの両方のビットマップ格納に置を 示す自ビットマップアドレスと、自コントローラのキャッシュ位置を示す自キャッシュアドレスと、創起2つの コントローラの他方のピットマップ格納位置を示す他ピットマップアドレスと、前記2つのコントローラの他方 のキャッシュ位置を示す他キャッシュアドレスを有する ことを特徴とする結束項10に記載のキャッシュメモリ 新舞力法に存する。また、この発明の請求項 1 2 に記載 の発明の要行は、前記ピットマップは、各セグメントに 対応するキャッシュのどの位置に有効なデータが存在す るかを示すマップであることを特徴とする請求項1 1 に 記載のキャッシュメモリ制御方法に存する。また、この 発明の緯次項13に記載の発明の要目は、前記ライトキ ャッシュは、各セグメントに対応するキャッシュメモリ であることを特徴とする請求項!2に記載のキャッシュ メモリ制御方法に存する。また、この免明の請求項 1-4 に記載の免明の整百は、前記ホストコンピュータからラ イトコマンドを受信した際に前記コントローラの一方 は、自己の前記セグメント管理情報を参照して自身がシ ーズしているセグメントにヒットしているかを確認する 工程と、セットしていなければ利用バスを介して前記2 つのコントローラの他方にシーズコマンドを返債してセ グメントをシーズする工程と、前記ホストコンピュータ

ているセグメントにヒットした場合に前起2つのコント いらセフメトドにセットに公司で開起とこのコント ローラの他がアン・ズコマンドを活体することかく、前 起ホストコンピュータからライトデータを交換して自己 の前記ライトキャッシュに移動しる基格所で置に対応し た自己の前記ライトキャッシュのデータおよび自己の と、自己の前記ライトキャッシュのデータおよび自己の

05 前紀ピットマップを前記2つのコントローラの他方のD MAコントローラモ介して前起2つのコントローラの他 方の前起ライトキャッシュと前起ピットマップに送ばす 10

る工程を有することを特徴とする請求項13に記載のキャッシュメモリ制御方法に存する。 [0022]

【90022】 【発明の漢語の階盤】 本発明は、2コントローラ構成 で、かつ、共有メモリのハードウェアを持たないディス クアレイ協語50において、ライトデータを開省から守 るための2番音に関して、メイトコントローラ (接着 すれば、相手コントローラ) のプロセッサに負荷をかけ ないことを特徴としている。 【0023】ライトデータはコントローラ内のメモリに

格約されるため、1コントローラ構成でライトキャッシュを使用すると、コントローラの故障によってライトデ ータを失ってしまう。---方、2コントローラ構成で、か つ、ライトデータを両方のコントローラに持っておけ ば、片方のコントローラが枚降しても、ライトデータが

失われることがない。このようにライトデータを両方の 下、本発明の実施の形盤を図面に基づいて詳細に説明す

【0024】 (第1の実施の形態) 図1は、本発明の第 1 の実施の形態に係るディスクアレイ装置 5 0 を説明するための機能プロック図である。本実施の形態のディス 30 クアレイ装置50は、ホストコンピュータ11とディス ク15, …, 18との間に2つのコントローラ21 1を有するものである。

75 【0025】ホストコンピュータ!」とコントローラ2 1.31は、例えば、SCS1 (スカジー: スモール・ コンピュータ・システム・インタフェース) パスのよう な汎用パス 1 2、 1 3 で接続されている。 【0 0 2 6】コントローラ2 1、 3 1 とディスク 1 5

…、18は、例えば、SCS1パスのような汎用パス1 4で接続されている。

【0027】コントローラ21は、マイクロプロセッサ 23上、リードまたはライトのデータをキャッシングし かつプロセッサの制御情報を記憶するメモリ27と、ホ

ストコンピュータ11との通信を誘導するインタフェ スチップ22と、ディスク15、…、18との通信を制 制するインタフェースチップ26と、自身のコントロー ラ21のメモリ27の任意アドレスのデータをメイトコ ントローラのメモリ27の任意アドレスに送録できるD を受信する工程と、前起コントローラの一方がシーズし 50 MAコントローラ24を有し、コントローラ21内の各

ディスクアレイ装置およびキャッシュメモリ制 有方法 特別2001-318766

チップは内部パス25で接続されている。

【0028】 同様に、コントローラ31は、マイクロブ ロセッサ33と、リードまたはライトのデータをキャッ シングしかつプロセッサの制御情報を記憶するメモリ3 7と、ホストコンピュータ11との通信を制御するイン タフェースチップ32と、ディスクしら、…. 18との 通信を制御するインタフェースチップ36と、自身のコ ントローラ31のメモリ37の任意アドレスのデータを メイトコントローラのメモリ37の任意アドレスに近ば できるDMAコントローラ34を有し、コントローラ3 1内の各チップは内部パス35で接続されている。

【0029】図2は、図1のメモリ27、37に記憶する情報の定量図である。本見明ではセグメントと呼ばれ る単位でキャッシュを管理しており、メモリ27、37 には、セグメント管理情報41、44と、ビットマップ 42、45と、ライトキャッシュ43、46が定義されている。セグメント管理情報41、44は、各セグメン トの状態を保持するもので、シーズの状態を示すシーズ フラグと、LUN(ロジカルユニット番号)と、LBA (ロジカルブロックアドレス) と、自コントローラのビ ットマップ格替位置を示す自ビットマップアドレスと、 自コントローラのキャッシュ位置を示すロキャッシュア ドレスと、メイトコントローラのビットマップ格制位置 を示す他ピットマップアドレスと、メイトコントローラ のキャッシュ位置を示す他キャッシュアドレスを有す る。ピットマップ42、45は、各セグメントに対応す るキャッシュのどの位置に有効なデータが存在するかを 示すマップで、0が無効、1が有効である。ライトキャ

ュメモリ自体である。 【0030】コントローラ21がホストコンピュータ1 【からライトコマンドを受信すると、セグメント管理情 報41を夢願して、自身がシーズしているセグメントに ヒットしているかを確認し、ヒットしていなければ武川 パス14を介してコントローラ3!にシーズコマンドを 透慮して、セグメントをシーズする。その核、ホストコ ンピュータ 1 1 からライトデータを受信してライトキャ **ッシュ43に格納し、当故格納位置に対応したビットで** ップ42を1に更新する。その後、ライトキャッシュ4 3のデータとピットマップ42をDMAコントローラ2 4、34**を**介してコントローラ31のメモリ37のライトキャッシュ46とピットマップ45に送ばする。ここ までの処理では、シーズコマンドの送受はによって、 ントローラ31のマイクロブロセッサ33が動作してい

シュ43、46は、各セグメントに対応するキャッシ

【0031】続けて、ライトコマンドを受付し、自身が シーズしているセグメントにヒットした場合、コントロ ーラ31にシーズコマンドを送ばすることなく、ホスト ャッシュ43に格納し、当故格納化器に対応したピット 50 爻はし(ステップS57:ライトデータ爻は)、それに

マップ42を1に更新する。その後、ライトキャッシュ 4 3のデータとピットマップ 4 2 をDMAコントローラ 2 4. 3 4 を介してコントローラ3 1 のライトキャッシ ュ46とピットマップ45に送信する。

【0032】以上説明したように第1の実施の影響によ れば、2回目以降のライトコマンドではコントローラ3 1のプロセッサ33に負荷をかけることなくライトコマ ンド処理が実行できるため、装置全体の性能が向上す

10 【0033】次にディスクアレイ装置50の動作(キャ ッシュメモリ制御方法)について説明する。なお、本実 施の彩蝶の動作について、ホストコンピュータ)I から ライトコマンドを受信し、それを2歳審さする例を示 。また、本実施の形盤では2回日以降のライトコマン ド処理に効果が変れるため、ライトコマンドを連続して

2 阿受信する場合で説明する。 【0034】図3は、本発明の第1の実施の形態に係る

キャッシュメモリ新郷方法におけるライト処理を説明す るためのフローチャートであって、汎用パス 1 4 に S C S I インタフェースを使用し、ライトコマンドを 2 回受

信した場合の処理の流れを示している。 【0035】コントローラ21は、ホストコンピュータ 1 1からライトコマンドを受信し (ステップS51:ラ イトコマンド受付)、自身のセグメント管理情報41を 25 参照してセグメントのヒットミスの特定を行う (ステッ

● 図し、モジメントのレジトミスの利用を行う」、バブ 「グ5 2 : ヒットミス特定」。 「回日のコマンドなので それがミスし、コントローラ3 I に見用バス I 4 を介し てシーズコマンド発行し(ステップ 5 5 3 : S C 5 I で シーズコマンド発行)、同時にそのパラメータとしてし UNとLBAを追信する。

【0036】それを受むしたコントローラ3 Lは、自身の売きセグメントをシーズして、セグメント管理情報4 4に、シーズ状態を示すフラグとLUNとLBAとシー ズしたキャッシュのアドレスとそれに対応するビットマップ42、45を格納するアドレスを登録して(ステッ

プS 5 4 : セグメント管理情報 4 4 に登録)、コントローラ21にシーズの完了報告を実行し(ステップS 5 5:SCSIでシーズ完了報告)、同時にそのレスポン ス情報としてキャッシュアドレスとピットマップ格納ア ドレスを送ばする。コントローラ21は自身のセグメン

40 ト代表は報4 1 に、シーズ状態を示すフラグとしびれと し、BAと自身がシーズしたキャッシュのアドレスとそれ に対応するピットマップ4 2、4 5 を格納するアドレス とメイトコントローラでシーズしたキャッシュのアドレ スとそれに対応するビットマップ42、45を格納する

アドレスを登録する(ステップSS6:セグメント管理 グドレスとは中すり、ベステンジリン・ピューター 情報44に登録)。 【0037】その後、ホストコンピューター1から自身 がシーズしているキャッシュアドレスにライトデータを

ディスクアレイ装置およびキャッシュメモリ制 御方法

からライトデータを受信して自己の前記ライトキャッシ ュに格納し当該格納位置に対応した自己の前記ピット

ップを有効状態に更新する工程と、自己の前記ライトキ

ャッシュのデータおよび自己の府記ピットマップを构起 2つのコントローラの他方のDMAコントローラを介し

て前記2つのコントローラの他方舗の前記ライトキャッシュおよび前記ピットマップに送信してライトコマンド

特別2001-318766

対応するビットマップ42を更新する(ステップS5 8:ピットマップ更新)。 【0038】何ライトデータを、DMAコントローラ2

、34を介してコントローラ31がシーズしているキ ャッシュアドレスにダイレクトに転送し(ステップSS 9 ; DMAでライトデータ転送)、同ビットマップ4 2、45€、同じくDMAコントローラ24、34を介 してコントローラ31のピットマップ45に転送し(ス テップS60:DMAでピットマップ転送)、これによって2重要さが完了し何時に1回日のライトコマンドが

元19 6。 (10039) 校けて、2回日のライトコマンドを受得すると(ステップS61:ライトコマンド受視)、自身のセグメント管理情報41を参照してセグメントのヒット ミスの料定を行い(ステップS62:ヒットミスギ 定)、それがヒットすると、1回目のようなシーズのや り取り無しにホストコンピュータ11から自身がシーズ しているキャッシュアドレスにライトデータを受信する ことができる(ステップS63:ライトデータ受信)。 【0040】その後、それに対応するビットマップ42 を更新して(ステップS64:ピットマップ更新)、同 ライトデータと同ピットマップ42、46を上記1回日 と同様にしてコントローラ31に伝送し(ステップS6 5:DMAでライトデータ転送、ステップS66:DM Aでピットマップ転送)、これによって2項書きが完了 し開時に 2回目のライトコマンドが完了する

【0041】これらの処理を比較すると、1回目のコマ ンド処理 (ステップS67) では、シーズコマンドのや り取りが行われて処理に時間がかかっているのに対し、 2回目以降のコマンド処理(ステップS 6 8)ではシー ズコマンドのやり取りがなく、処理時間が短縮され、か つ、コントローラ31のプロセッサに負債がかかってい ないことがわかる。

【0042】図4は、本発明の第1の実施の形態に係る ディスクアレイ装置50の各コントローラ21、31に おけるライト処理を説明するためのフローチャートであって、コントローラ21とコントローラ31のライトコ マンド処理のフローチャートを示している。

【0043】図4を幕照してそれぞれのプロセッサがどのような処理を行うかを説明する。ステップS71(コ ントローラ21の処理)の実行後、コントローラ21 は、ホストコンピュータ11からライトコマンドを受信 すると(ステップS72:ホストコンピュータ11から ライトコマンド受(は) 、自身がシーズしているセグメン トにヒットしているかどうかを判断する(ステップS7 3:シーズ中セグメントにヒット?)。

【0044】ここで1回目ならミスするので(ステップ 873のNo)、SCS1にてコントローラ31ヘセグ メントシーズコマンドを発行する (ステップS74:S C S 1 にてコントローラ31ヘセグメントシーズコマン

F Office

15

【0045】コントローラ31はそのシーズコマンドを 受信(ステップS82:コントローラ31の処理→ステ ップS83:SCSIにてコントローラ21からセグメ ントシーズコマンド受信)し、自身のセグメント管理情 報44を更新して(ステップS84:セグメントを理像 報44の更新)、コントローラ21にセグメントシーズ 成功の適知を送信する(ステップS85;SC51にて コントローラ21ヘセグメントシーズ成功の通知を送

【0046】コントローラ2:はその通知を受信し(ス テップS75:SCSIにてコントローラ31からセグ メントシーズ成功の通知を受信)、自身のセグメント管 アン・・・ スルリン MA (ことの)、 ロガヤン・・ン 入 関情報 4 1 の更新) 、 (ステップS 7 6 : セグメント登 単情報 4 1 の更新) 、 ホストコンピュータ 1 1 からライ トデータを受信し (ステップS 7 7 : ホストコンピュー タ11からライトデータ受信)、それに対応してピット マップ42を更新する(ステップS78:ビットマップ 42の更新)

【0047】その後、DMAにてコントローラ31にラ ZO イトデータを送信し(ステップS79:DMAにてコントローラ31ヘライトデータを送保)、続いてビットマ ップ42を送供し(ステップS80:DMAにてコント ローラ31ヘビットマップ42を送信)、これらの完了 をもってホストコンピュータ11へ完了報告を行う(ス

25 テップ581:ホストコンピュータ11へ充了報告)。 【0048】2回日以降は、シーズ中セグメントのヒッ ト判定(ステップS73:シーズ中セグメントにヒット (2) においてヒットとなり (ステップS 7 3 の Y e
 (30 s)、シーズ関連の処理 (ステップS 7 4 : S C S I に

てコントローラ3 | ヘセグメントシーズコマンド発行。 ステップS75:SCS | にてコントローラ3 | からセ グメントシーズ成功の通知を受信、ステップS76:セ グメント管理情報 4 1 の更新) をスキップする。これに より、コントローラ3 1 の処理(ステップS 8 2)が行 われなくなるので、マイクロプロセッサ33の負債がか

【0049】以上説明したように第1の実施の形態によ れば、メイトコントローラのセグメントをシーズしてそ のキャッシュアドレスとピットマップ42、45を保持 いイマンスプレンスでにファイフンスで、マロミル することにより、2回日以降のライトコマンド処理にお いてはメイトコントローラのプロセッサに負荷をかける ことなく実行できるようになり、その結果、装置全体の 性能の向上を図ることができるようになるといった効果

を分する。 【0050】 (第2の実施の形盤) 以下、本発明の第2 の実施の形態を成明する。なお、上紀第1の実施の形態 において数に記述したものと同一の部分については、同 ·存付を付し、仮投した説明は省略する

【0051】本発明の第2の実施の形態は、上記第1の

2003 04 24 10:41

2003 04 24 10 41

実施の影響の基本構成に加大で、コントローラ2:1、3 1 散を増やすこともできる構成を備えている点に時度を 有している。 特体的には、シーズコマンドを返出するだ。 のの、例えば、SCS 1 バスのような表別バスと、自身 のメモリ2:7、3 7 のデータを任息のコントローラ2 1、3 1 のメモリ2:7、3 7 にダイレクトに近ばできる DM A コントローラ (不図示)を 有している。

DMAコントローラ (本図ぶ)を打している。 (10052) 効理の機能は、比較期1の実施の形態の2 コントローラ解表の場合と基本的に同じであり、2 電点 をしようとするコントローラ21、3 1に対して1回口 はシーズコットを活送してビグメントをシーズル。2 回目以降はシーズ型理集して2番号を必折する。 (100531かお、本例明上記令資金の形態に限いさ 対す、本規則の採集型の規則において、各項条の形態 に対して変更され得ることは明らかである。また上記機

■は無止之でいっています。 成部所の数、位置、形状寺は上紀実施のお憩に敬定され ず、本発明を実施する上で好遊な数、位置、形状等にす ることができる。また、各図において、同一構成要素に は同一符号を付している。 [0054]

【発明の効果】 本発明は以上のように構成されているの 1 実明の効果」 本発明は以上のように構成されているの 、メイトコントローラのセグメントをシーズしての キャッシュアドレスとピットマップを保持することによ り、2回日以際のライトコマンド処理においてはメイト コントローラのプロセッサに負荷をかけることが、気持 できるようになり、その4原、装置合体の性能の向上を 図ることができるようになるといった効果を挙する。 (WWW.OPERCED) 【団面の簡単な故明】

【図1】 本発明の第1の実施の形盤に係るディスクアレ

イ装置を設断するための機能プロック図である。 【図2】図1のメモリに記憶する情報の定義図である。 【図3】 本発明の第1の実施の形態に係るキャッシュメ モリ朝御方法におけるライト処理を説明するためのプロ

ーチャートである。 【図4】本発明の第1の実施の形態に係るディスクアレ イ装置の各コントローラにおけるライト処理を説明する ためのフローチャートである。

【図5】 節!従来技術のキャッシュメモリ制御方式を説

明するための機能プロック図である。 【図6】 第2従来技術の主記総装器の制御方式を説明す るための機能プロック図である。 【図7】 第3従来技術のディスクアレイ装置を放明する ための機能プロック図である。

[行けの説明]

11…ホストコンピュータ 12. 13. 14…摂用パス 15. …. 18…ディスク 21. 31…コントローラ

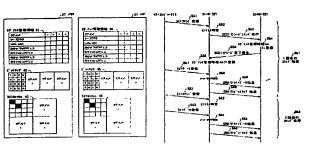
22、26、32、36…インタフェースチップ 23、33…マイクロプロセッサ 24、34…DMAコントローラ

25.35…内部パス 27.37…メモリ

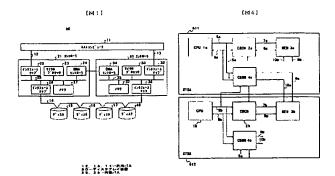
25 41、44…セグメント管理情報 42、45…ビットマップ 43、46…ライトキャッシュ

50…ディスクアレイ協選

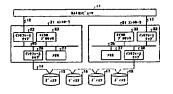
[图2]



### ディスクアレイ装置およびキャッシュメモリ制 御方法



[847]

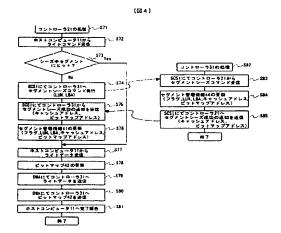


12, 13, 14- KR/G

2003 04 24 10:41

### ディスクアレイ装置およびキャッシュメモリ制 御方法

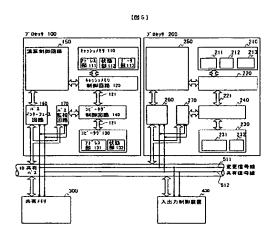
特別2001-318766



- 11 -

### ディスクアレイ装置およびキャッシュメモリ制 御方法

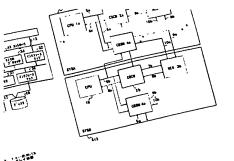
特別2001-318766



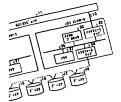
- 10 -

フロントページの続き			
(\$1) fat. Ct. 7	識別記り	FI	テーマコード(参考)
G 0 6 F 12/16	310	G 0 6 F 12/16	310)
	320		3201.
G 1 1 B 19/02	5 0 t	G 1 1 B 19/02	501F

2003 04 24 10-41 2003 04 24 10.41



18471



**特別2001-318766** ティスクアレイ装置およびキャッシュメモリ制 制方法

CO 6 E 12/16 C 1 1 B 19/01 (51) ln1.Cl. 1 G 0 6 F 12/16 Q 1 1 B 19/02

- Data that is accessed normally with some locality of reference will use partial track mode staging. This is the default mode.
- Data that is not a regular format, or where the history of access indicates that a full stage is required, will set the full track mode.
- The adaptive caching mode data is stored on disk and is reloaded at IML

### aniontial roads

Cache space is released according to Least Recently Used (LRU) algorithms. Space in the cache used for sequential data is freed up quicker than other cache or record data. The ESS will continue to pre-stage sequential tracks when the last few tracks in a sequential staging from a sequential staging

Stage requests for sequential operations can be performed in parallel on the RAID array, giving the ESS its high sequential throughput characteristic. Parallel operations can take the bace because the logical data tracks are striped across the physical data disks in the RAID array.

# 3.28 NVS and write operations

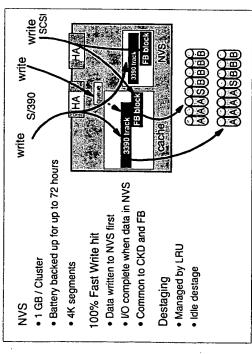


Figure 3-32 NVS - write

As Figure 3-32 illustrates, at any moment there are always two secured copies of any update into the ESS.

## 3.28.1 Write operations

Data written to an ESS is almost 100% last write hits. A tast write hit occurs when the write VO operation completes as soon as the data is in the ESS cache and non-volatile storage (NVS). The benefit of this is very fast write operations.

### ast write

Data received by the host adapter is transferred first to the NVS and a copy held in the host adapter buffer. The host is notified that the I/O operation is complete as soon as the data is in. NVS. The host adapter, once the NVS transfer is complete, the transfers the data to the cache.

Cache.
The data remains in the cache and NVS until it is destaged. Destage is triggered by cache またれい and NVS usage thresholds.

### 8.2 NVS

The NVS size is 2 GB (1 GB per cluster). The NVS is protected by a battery. The battery will power the NVS for up to 72 hours following a total power failure.

10

### VVS LRU

NVS is managed by a Least Recently Used (LRU) algorithm. The ESS attempts to keep free space in the NVS by anticipatory destaging of tracks when the space used in NVS exceeds a threshold. In addition, if the ESS is ide for a period of time, an idle destage function will destage sund; affine about 5 minutes, all tracks will be destaged.

Both cache and NVS operate on LRU lists. Typically space in the cache occupied by sequential data is released earlier than space occupied by data that is likely to be re-referenced. Sequential data in the NVS is destaged ahead of random data.

When destaging tracks, the ESS attempts to destage all the tracks that would make up a RAID stripe, minimizing the RAID-related activities in the SSA adapter.

### IVS location

NVS for cluster 1 is located physically in I/O drawer of cluster 2, and vice versa. This ensures that we always have one good copy of data, should we have a failure in one cluster.

See 3.8, "Cluster operation: fallover/failback" on page 60 for more information.

### This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

### **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

☐ BLACK BORDERS
☑ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
☐ FADED TEXT OR DRAWING
☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
☐ SKEWED/SLANTED IMAGES
☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
☐ GRAY SCALE DOCUMENTS
☐ LINES OR MARKS ON ORIGINAL DOCUMENT
☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
□ OTHER:

### IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.